

DOCUMENT RESUME

ED 216 357

CS 206 870

AUTHOR Bauer, Barbara Ann
TITLE A Study of the Reliabilities and the Cost-Efficiencies of Three Methods of Assessment for Writing Ability.
PUB DATE 81
NOTE 22p.; Research prepared at the University of Illinois.
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cost Effectiveness; Expository Writing; *Holistic Evaluation; Secondary Education; *Test Reliability; *Writing Evaluation; Writing Instruction; *Writing Research
IDENTIFIERS *Analytic Scoring; Interrater Reliability; *Primary Trait Scoring

ABSTRACT

To compare the relative reliable uses and cost effectiveness of the analytic, the holistic, and the primary trait scoring methods, an inquiry was conducted in which a group of raters scored a large number of secondary school students' essays according to each of the scoring methods. Raters were nine graduate students in English who were trained in the use of each of the three scoring methods. The raters were then assigned to one of three groups, each using one scoring method exclusively. Essays scored were drawn from those generated as part of the 1973-74 writing assessment of the National Assessment of Educational Progress. Data produced in the study were analyzed for the interrater and intrarater reliabilities; the comparative rater reliabilities; the effects and the interactions that the methods, the raters, and the essays had on essay scores, as well as on one another; and the cost effectiveness in terms of training and grading times for each of the three scoring methods. Results showed that the analytic scoring method was the most reliable and the holistic scoring method the most cost effective in assessing large numbers of student essays. (FL)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

+ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

A STUDY OF THE RELIABILITIES
AND THE COST-EFFICIENCIES OF THREE METHODS
OF ASSESSMENT FOR WRITING ABILITY

By Barbara Ann Bauer

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Barbara Ann Bauer

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Dr. Barbara Ann Bauer
Graduate Student
331 Sherman Hall
University of Illinois
909 South Fifth Street
Champaign, Il. 61820
Telephone: (217) 332-4703

A STUDY OF THE RELIABILITIES
AND THE COST-EFFICIENCIES OF THREE METHODS
OF ASSESSMENT FOR WRITING ABILITY

Barbara Ann Bauer
University of Illinois

Many English teachers will attest to the fact that assessing achievement in the performance of writing ability is often an unreliable and time-consuming task, especially when a large number of student essays is being graded. These two factors are ever-present in the grading process; but the extent to which they become problematic to this process depends largely upon the degree to which different raters assign different grades to the same essay. This phenomenon, in turn, is influenced by the particular conceptualization of writing ability that the raters adhere to while grading essays. The particular conceptualization that raters do adhere to is, then, reflected in the type of grading method they use. While "most of these (methods) may be reduced to variants either of the impressionistic principle or of the analytic principle ..." (Burt, in Cast, 1939, p. 258), other types of methods do exist; one such method is the primary trait scoring system. Though this is the case, much attention in the past has been given to the problem of determining whether the analytic or the impressionistic/holistic method is more reliable and/or cost-efficient in grading essays, and little to no attention has been given to comparing the reliabilities and the cost-efficiencies of these two methods with those of the primary trait method. This study makes such a comparison, where the analytic, the holistic, and the primary trait scoring methods each conceptualize achievement in writing ability differently.

REVIEW OF PREVIOUS RESEARCH

Previous research concerned with the problem of obtaining rater-reliability in the assessment of writing ability in student essays is of two main types: that research which helped to develop the analytic, the holistic, and the primary trait conceptualizations of writing ability into three valid methods of assessment; and that research which attempted to compare the rater-reliabilities of these established methods, one with the other.

Developing Conceptualizations into Methods of Assessment for Writing Ability

The analytic scoring principle conceives of achievement in writing ability as the ability to produce a piece of written discourse having certain characteristics or "elements" which contribute to the composition paper as a whole. This conceptualization of writing ability is best exemplified in the Diederich Expository Scale, which was developed through several decades of research by Paul B. Diederich and his associates. This "scale" was theoretically established through the work of Diederich, French, and Carlton (1961) in which it was found that the reader reliability among 60 readers of 300 essays was .31, and that different readers assigned different grades to the essays for a variety of reasons. These reasons were factor-analyzed into five main types: 1) ideas; 2) form; 3) flavor; 4) mechanics; and 5) wording. With this theoretical framework, the scale was modified by Diederich (1966; 1974) so as to improve reader-reliability and to encourage its practical use in the grading of student essays. These modifications included the following: "a separate rating on the four main components (i.e., ideas, organization, wording, and flavor) ... and ... a rating on 'handwriting neatness' as well as on usage and sentence structure, punctuation, and spelling" (Diederich, 1966, p. 444); the allocation of ten points for the ideas and organization components, and

just five points for the other six components; and the incorporation of a preliminary "first impression" reading of the papers. As a result, the revised edition of the Diederich Expository Scale was found to be both practical and highly reliable; with regard to the latter, Diederich (1974) found a reader-reliability of .80 when an essay was graded by two trained readers according to this scale.

The holistic scoring principle conceives of achievement in writing ability as the ability to produce a piece of written discourse having a unified whole to which the various characteristics, together, constitute the composition paper as a whole. This conceptualization of writing ability is best exemplified in the 4-point holistic scoring method which is currently being used to score the English Composition Test by the College Entrance Examination Board. This particular holistic scoring method was primarily established through the work of Godshalk, Swineford, and Coffman (1966). Although Godshalk, et al. (1966) obtained an estimated reader-reliability of .921 when 25 experienced readers graded 646 essays on a three-point holistic scale (with five readings per essay), there was evidence that some readers used the "safe procedure" of awarding "2's" to those essays whose over-all quality was in doubt. To rectify this problem, Godshalk, et al. (1966) conducted a field study in which 533 essays from the original study were re-read by 146 readers on both a three-point and a four-point holistic scale; the four-point holistic scale was found to be more highly favored and more reliable than the three-point scale. This four-point holistic scale was somewhat modified by the College Board so that it could be practically applied to the scoring of the essay in the 1977 edition of the English Composition Test.¹ Such modifications included the independent scoring of each essay on a four-point scale with no mid-point; the incorporation

of two types of judgment into the grading process; and the setting of the standards for grading by the readers in an intensive training session. Thus, the four-point holistic scale has been found to be both practical and reliable.

The primary trait scoring principle conceives of achievement in writing ability as the ability to produce a piece of written discourse having all of the traits or "features" that are necessary in accomplishing its aims, and these features include a particular blend of audience, speaker role, purpose, and subject. This conceptualization is primarily exemplified in the Primary Trait scoring system which was developed by the National Assessment of Educational Progress, and designed to be used in the 1973-74 second round of assessment in writing ability.² Because National Assessment wanted a scoring system that was replicable in future assessments, more clearly defined the differences between the score points assigned to the essays, and provided more specific information about young people's abilities to produce different kinds of writing, it developed a scoring system which was based on a three-part rhetorical model. From this model, which included a consideration of explanatory, persuasive, and expository writing, National Assessment devised various types of writing exercises that were to be judged according to narrowly defined situations of the writer. To assess achievement in the defined skill or trait contained in the particular exercise, a corresponding scoring guide was also developed with four basic levels of competency; each level of competency is represented by a specific score point value, with a score-point of "1" representing the lowest level and a score-point of "4", the highest level. Using this scoring system in the second round of assessment in writing, National Assessment found that the percentages of reader agreement between two readers' scores on each essay, and before reconciliation with a third reader, were generally high.

Comparing the Rater-Reliabilities of Various Methods of Assessment for Writing Ability

The research of Cast (1939; 1940), and more recently, that of Britton, Martin, and Rosen (1966), Follman and Anderson (1967), Marzano (1975), and Winters (1978) has attempted to compare the reliabilities of raters' uses of those established scoring methods as delineated above. With the exception of the Britton, et al. (1966) study, these studies revealed that, although the differences were small, the analytic scale(s) tended to be more reliably used than the holistic scale(s), even when raters had similar experiential backgrounds and have been given adequate training in the use of both types of scoring methods. But, it was also revealed that no one of these studies had considered the reliable uses of the primary trait scoring method in relation to the reliable uses of either the analytic or the holistic method of assessment. With this in mind, this study did consider the relative reliable uses of the analytic, the holistic, and the primary trait scoring methods by addressing the following questions:

1. What is the comparative rater-reliability of the three methods?
2. In what method is the frequency of drift most apparent--i.e., in what method is there the least amount of intra-rater reliability?
3. Which method appears to require the least amount of training time?
4. Which method appears to require the least amount of grading time per essay?
5. Which method would appear to be the most cost-efficient in the assessment of large numbers of student essays?

METHOD OVERVIEW

To compare the relative reliable uses of the analytic, the holistic, and the primary trait scoring methods, an empirical inquiry was conducted in which a group of raters scored a large number of student essays according to each of these three scoring methods.

The Writing Sample

The sample of student essays used in this study was secured from among 737 seventeen-year-olds' responses to a single writing exercise in the narrative-descriptive mode of discourse, as part of the 1973-74 second round of assessment in writing for the National Assessment of Educational Progress. From this national sub-sample of essays, which were computer counter-transcriptions of the original handwritten responses, a smaller sample of 120 essays was chosen, and this smaller sample represented the four main levels of competency as reflected in the primary trait scoring guide that had originally been used to score the essays in the 1973-74 National Assessment study.

Because this smaller sample of 120 essays was used for both training and grading purposes, it was sub-divided into two groups of essays. The first group of essays was used for training purposes, consisting of a "pool" of 20 essays with a random selection of five essays from each of the four levels of competency. The second group of essays was used for grading purposes, and although it was to have consisted of the remaining 100 essays, two of these essays had to be omitted from this group, thus bringing the number down to 98 essays. But, since this group of 98 essays was to be used in determining the intra-rater reliabilities of the three methods of assessment, 20 of these essays were selected for duplication purposes. Thus, the second group of essays used in the grading process totaled 118; that is, 98 common essays and 20 duplicated essays.

The Raters

The raters who read and scored the sample of essays were selected from among 70 teaching assistants enrolled in the graduate division of the English Department at the Urbana-Champaign campus of the University of Illinois during the 1980 Fall Semester. Consisting of nine graduate teaching assistants--i.e., five women and four men, none of these raters had previous systematic training in the use of any one of the three methods of assessment ascribed to here. These nine raters were first asked to participate in the grading of 118 essays, and then divided into three groups of three raters, where each group of raters was assigned to one of the three scoring methods. Thus, the raters were grouped as follows: the group of three raters who were assigned to using the analytic scoring method consisted of two women and one man; the group of three raters who were assigned to using the holistic scoring method consisted of three men; and the group of three raters who were assigned to using the primary trait scoring method consisted of three women.

The Procedures Used in Obtaining the Data

The procedures used in obtaining the data for this study were three-fold: the preliminary preparations; the training sessions; and the grading sessions.

1. Prior to the training and the grading sessions, some preliminary preparations were completed; they included the design and development of the score sheets that were used by the raters to score the essays, and the preparation of the essays themselves for the training and grading sessions. Because of the large number of essays to be graded, special score sheets were designed for use in this study. These score sheets adapted the original rubrics of the analytic, the holistic, and the primary trait scoring methods into a type of

multiple-choice test item format so that they could be read by an op-scan computer. Together with the reformulated scoring rubrics, and the special directions to the raters printed on them, three sets of score sheets were developed; one set was printed up with the analytic scoring rubric, a second set with the holistic scoring rubric, and a third set with the primary trait scoring rubric.

In addition, the sample of essays was prepared for its use in the training and the grading sessions. After separating those essays that had been selected for training and grading purposes from the rest of the print-out, the 118 essays were individually attached to separate sheets of plain white paper. From these essays, 40 of them were duplicated by means of xeroxing; that is, six copies of the 20 essays earmarked for training purposes were xeroxed, and one copy of the 20 essays that were to serve in determining the intra-rater reliabilities of each method was also xeroxed.

The score sheets were then assembled with the prepared essays and placed in their respective groups. The six sets of 20 essays used for training purposes were stacked together so as to form a "pool" of essays. For those essays used in the grading sessions, each of the 118 essays was paper-clipped to three score sheets, and divided into three piles--i.e., 35, 38, and 45 essays respectively; this process was repeated three times to accommodate for the differences in the three types of scoring methods.

2. The training sessions were conducted prior to the actual experiment so as to provide the raters with intensive training in the scoring of a large number of essays according to the analytic, the holistic, and the primary trait scoring principles. Because three different groups of raters were to grade these essays according to the three different scoring methods, three separate training sessions were conducted, each one being conducted on a separate occasion; that is, the first training session was used to train one group of raters to

grade the essays analytically, the second was used to train another group of raters to grade the essays holistically, and the third was used to train the last group of raters to grade the essays according to a specific primary trait. Each of these three training sessions was conducted at a central location on the Urbana-Champaign campus of the University of Illinois by Professor Alan C. Purves. In addition, this investigator presided over each of these three training sessions in an effort to monitor the amount of time needed to train each group of raters so as to help determine the cost-efficiency of each method.

Although each of these three training sessions followed the same general procedures, certain variations in their implementation did exist within each session. Such variations included a slight modification of a scoring rubric to accommodate the analytic group of raters--e.g., the omission of the "hand-writing" characteristic from use in this study since the essays were computer counter-transcriptions, as well as slight deviations in the ways that the raters were trained to use the holistic and the primary trait scoring methods for this study as compared with the conventions of the Educational Testing Services and National Assessment respectively. Be this as it may, however, each training session served to aid in enhancing each group of raters' abilities to understand and internalize the respective scoring rubric among themselves, and then to apply that rubric to the grading of various essays selected from the "pool" of 20 training papers to the point where between-rater agreement was attained within that group.

3. The grading sessions were those times during which the raters read and graded the 118 student essays according to the analytic, the holistic, and the primary trait scoring methods. Because of this, three separate grading sessions were conducted, each one following its respective training session; that is, the first grading session was for the analytic scoring of the essays,

the second was for the holistic scoring of the essays, and the third was for the scoring of the essays according to a specific primary trait. Each of these three grading sessions was conducted in the same central location as the training sessions. Because consistency spot-checks were to be incorporated into each of the three grading sessions, Professor Alan C. Purves served as a Table Leader. As with the training sessions, this investigator presided over each of the three grading sessions in an attempt to monitor the amount of time needed to grade the essays so as to help determine the cost-efficiency of each method. The same general procedures were followed by each group of raters in each of the three grading sessions, and they included the following instructions:

- a. work as quickly as possible, but take a break when fatigue became a factor; scheduled breaks were incorporated into each grading session for this purpose.
- b. when scoring the essays, remove the top score sheet, fill out both sides, read and score the essay; and then place the marked score sheets on one pile, and the essays with the remaining attached score sheets on another pile.
- c. go on to score the second and third piles of prepared essays after the first pile of essays have been scored; each pile of prepared essays is to be circulated counter-clockwise until each of the three piles of essays has been scored by each of the three raters.

RESULTS

The results of this study included an analysis of the inter- and intra-rater reliabilities; the comparative rater-reliabilities; the effects and the interactions that the methods, the raters, and the essays had on the essay

scores as well as on one another; and the cost-efficiencies in terms of training and grading times for the analytic, the holistic, and the primary trait scoring methods.

The inter-rater reliabilities of the three scoring methods were obtained through an analysis of variance, which is based on Snedecor's intra-class correlation formula. The reliability estimates for the individual and average ratings for each of the three scoring methods are presented in Table 1.

Table 1
Inter-Rater Reliability Correlation Coefficients Among the
Scoring Methods

Method	Inter-Rater Reliability	
	Individual	Average
Analytic	.874	.954
Holistic	.811	.928
Primary Trait	.633	.838

As this table indicates, the reliability estimates were highest for the analytic method, and lowest for the primary trait method.

In an effort to determine whether or not the relative inter-rater reliabilities for the three scoring methods were significantly different from one another, a z-test for independent samples was performed using Fisher's Z transformations. With alpha (α) set at the .01 level of significance, this test showed that the differences between the reliability estimates of the analytic and the holistic

scoring methods were not significant for either the individual ratings (1.51) or the average ratings (1.51); but, a significant difference was evident between the reliability estimates of the analytic and the primary trait scoring methods for both the individual ratings (4.01) and the average ratings (4.41). The differences between the reliability estimates of the holistic and the primary trait scoring methods were not significant for the individual ratings (2.56), but significant for the average ratings (2.90).

In addition, the causes for the variations in the scores assigned to the essays by the raters using each of the three scoring methods was investigated by performing an analysis of variance using a 91×3 repeated measures design with z-score transformations. With alpha (α) set at the .01 level of significance, a significant main effect was found for the essays factor as well as for the interaction of the essays and the methods; although the F-ratio for the former (28.07) was clearly more significant than the F-ratio for the latter (1.44). No significant effect was found for the method factor ($F=0$).

The intra-rater reliabilities of the three scoring methods were obtained through Pearson's product-moment correlation coefficients by pairing each rater's scores on the 19 duplicate essays with each one's scores on the corresponding 19 "common" essays. The individual, the average, and the composite intra-rater reliability correlation coefficients are presented in Table 2.

Table 2

Intra-Rater Reliability Correlation Coefficients Among the Raters
Within Their Respective Scoring Methods

Method/Rater	Intra-Rater Reliability		
	Individual	Average	Composite
Analytic		.900	.897
Rater 1	.884		
Rater 2	.895		
Rater 3	.920		
Holistic		.785	.733
Rater 4	.849		
Rater 5	.773		
Rater 6	.733		
Primary Trait		.836	.820
Rater 7	.851		
Rater 8	.844		
Rater 9	.812		

As can be seen from this table, the degree of consistency that each individual rater had obtained within him/herself when asked to grade the same essay twice was highest for Rater 3 and lowest for Rater 6. It is also worth noting that the individual correlation coefficients for Raters 1, 2, and 3, all of whom used the analytic method, were higher than those for any of the other six raters.

The average and composite correlation coefficients were highest for the analytic scoring method, followed by the primary trait scoring method, and finally the holistic scoring method.

The cost-efficiencies of the analytic, the holistic, and the primary trait scoring methods were seen in terms of the time required to train each group of raters to use each of the scoring methods, and the time required for those raters to apply these methods to a large number of essays. In considering the former, the training time in minutes for each of the three scoring methods may be found in Table 3.

Table 3
Time Required to Train the Raters to Use Their Respective
Scoring Methods

Method	Training Time in Minutes
Analytic	114
Holistic	52
Primary Trait	44

As Table 3 indicates, the time required to train a group of raters to use its respective scoring method was longest for the raters who were assigned to the analytic method; this was followed by that group of raters who were assigned to the holistic method, and then by those raters who were assigned to the primary trait scoring method, which required the least amount of training time.

In addition, and perhaps in conjunction with the required training times for each of the three methods, the cost-efficiency of each of these methods was considered in terms of the time required for a group of raters to grade a large sample of essays according to a given scoring method. This is most conveniently seen in the average time required to grade a single essay within each of the three scoring methods, as presented in Table 4.

Table 4
Average Time Required to Grade a Single Essay Within
Each of the Scoring Methods

Method	Average Grading Time per Essay in Minutes
Analytic	4.14
Holistic	.96
Primary Trait	1.08

As Table 4 indicates, the average time needed to grade a single essay was longest for the analytic scoring method, and shortest for the holistic scoring method; although the differences in time between the holistic and the primary trait scoring methods appear to be minimal.

CONCLUSION

The conclusions that were drawn from this study's findings are two-fold: that which determines whether the analytic, the holistic, or the primary trait scoring methods is the most reliable in assessing achievement in writing ability in student essays; and that which determines the most cost-efficient scoring

method in the assessment of large numbers of essays.

In determining which of the three scoring methods is the most reliable, one needs to consider both the comparative reliabilities of the three methods and the degree of intra-rater reliability that is inherent within each method. When the three scoring methods are viewed in terms of their comparative rater-reliabilities, the analytic scoring method emerges as having a higher degree of reliability than the holistic and the primary trait scoring methods in terms of its inter-rater reliability, and this may be explained by the specificity of the criteria on which the scoring method is based; and the amount of training time that that group of raters used to come to an agreement with one another on the conceptualization of writing ability as reflected in that scoring method; and then the application of that conceptualization to the student essays. In terms of the degree of intra-rater reliability that is inherent in each of the three scoring methods, the analytic method also appears to have the highest degree of within-rater consistency and the holistic method, the least; this, too, seems to be influenced by the degree of structure and specificity of the scoring method. This is to say, because the analytic scoring method is highly structured and specific in terms of its grading criteria, the group of raters who did use it were not only able to agree with one another on its meaning and application to student essays, but also able to internalize its criteria within themselves and to use it consistently. With this in mind, the raters who used the holistic scoring method seem to have failed on the latter point, and the raters who used the primary trait scoring method seem to have failed on the former.

In determining which of the three scoring methods is the most cost-efficient, one needs to consider the amount of training time required by each of the three scoring methods, and the average time needed to score a single essay according

to each method. When the three scoring methods are considered in terms of each one's required training time, the analytic method appears to require the longest amount of time, followed by the holistic method, and then by the primary trait method. While these differences may be attributed, in part, to the nature of the scoring methods, they may also be attributed to the point at which it was felt that the raters in each group had attained a "moderate" degree of consistency among themselves. In terms of the average time required by any one rater within his/her group to score a single essay according to each of the three scoring methods, the analytic method also required the longest amount of grading time, and although the differences appear to be minimal, less grading time per essay was required when the holistic method was used than when the primary trait method was used; these differential grading times seem to be influenced by the nature of the scoring method that was used. On the basis of these two factors, and that of the degree of attained rater-reliability in the grading session, it may be seen that the analytic method fails on the first two counts and the primary trait method fails on the latter; but the holistic method emerges as being the most cost-efficient in the assessment of a large number of essays.

Thus, on the basis of this study's findings, it was concluded that the analytic scoring method was the most reliable and the holistic scoring method was the most cost-efficient when used to assess achievement in writing ability in large numbers of student essays.

NOTES

1. College Entrance Examination Board. Announcing the English composition test with essay. Princeton, N.J.: Educational Testing Service, 1977.
2. National Assessment of Educational Progress. The second national assessment of writing--New and reassessed exercises with technical information and data (Report No. 05-W-20). Denver, Col.: Education Commission of the States, May 1978.

References

- Britton, J.N., Martin, N.C., & Rosen, H. Multiple marking of compositions--An account of an experiment. London: Her Majesty's Stationary Office, 1966.
- Burt, C. Report on junior county scholarship examinations. In E.M.D. Cast, The efficiency of different methods of marking English composition: Part i. British Journal of Educational Psychology, 1939, 9(3), 257-269.
- Cast, E.M.D. The efficiency of different methods of marking English composition: Part i. British Journal of Educational Psychology, 1939, 9(3), 257-269.
- Cast, E.M.D. The efficiency of different methods of marking English composition: Part ii. British Journal of Educational Psychology, 1940, 10(1), 49-58.
- Diederich, P.B. Measuring growth in English. Urbana, Ill.: National Council of Teachers of English, 1974.
- Diederich, P.B., French, J.W., & Carlton, S.T. Factors in judgments of writing ability (Research Bulletin RB 61-15). Princeton, N.J.: Educational Testing Services, August 1961.
- Follman, J.C., & Anderson, J.A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 1(2), 190-200.
- Godshalk, F.I., Swineford, F., & Coffman, W.E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.
- Marzano, R.J. On the validity of analytic ratings. Denver, Col.: University of Colorado, 1975. (ERIC Document Reproduction Service No. ED 112 412)
- Pearson, R. The test fails as an entrance examination. In, Should the general composition test be continued: College Board Review, 1955, (25), 2-10.
- Peixotto, H.F. The relationship of the college board examination scores and reading scores for college freshmen. Journal of Applied Psychology, 1946, 30, 406-411.
- Remondino, C. A factorial analysis of the evaluation of scholastic compositions in the mother tongue. British Journal of Educational Psychology, 1959, 29(3), 242-251.
- Smith, V.H. Measuring teacher judgment in the evaluation of written composition. Research in the Teaching of English, 1969, 3(2), 181-195.

Traxler, A.E., and Anderson, H.A. The reliability of an essay test in English. The School Review, 1935, 43(7), 534-539.

Winters, L. The effects of differing response criteria on the assessment of writing competence. In E. Baker & E. Quellmalz (Dir's.), Studies in measurement and methodology--Work unit 1: Design and use of tests (OB-NIE-G-78-0213). Los Angeles: University of CA, Center for the Study of Evaluation, November 1978.

Wiseman, S. The marking of English composition in grammar school selection. British Journal of Educational Psychology, 1949, 19(3), 200-209.

Wiseman, S. Symposium: The use of essays in selection at 11+ III.--Reliability and validity. British Journal of Educational Psychology, 1956, 26(3), 172-179.